Mortality Matrix

ML for Predictive Health Care

TOC

INTRODUCTION

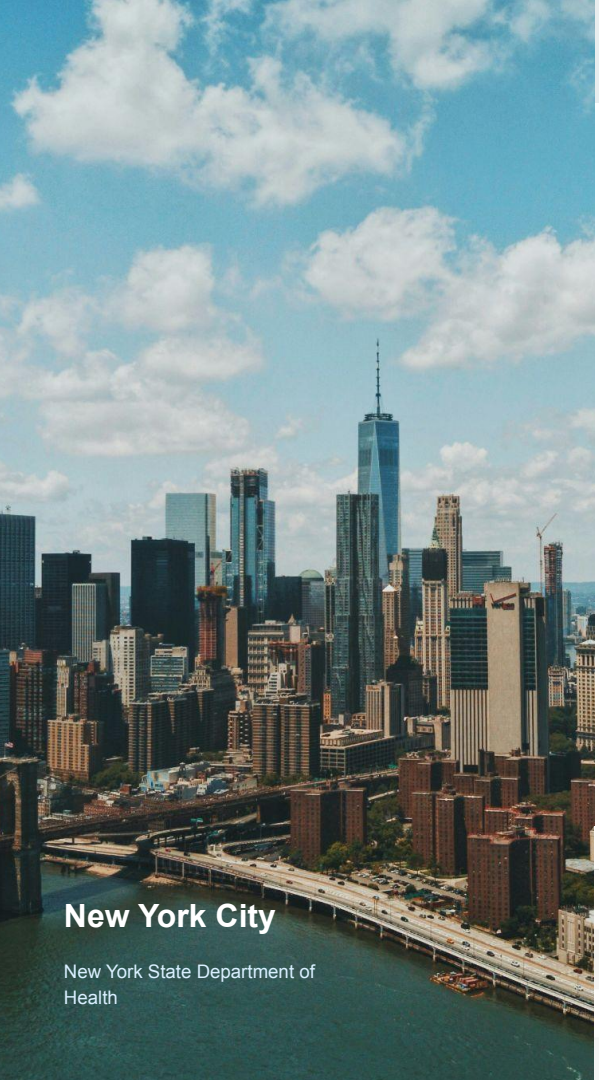
In the evolving field of healthcare analytics, the capability to accurately predict patient outcomes is essential for effective clinical decision-making and resource management. This project focuses on the application of advanced machine learning techniques to predict the All Patient Refined (APR) Risk of Mortality. By harnessing a rich dataset that includes patient demographics, clinical diagnoses, procedures, and outcomes, we aim to develop models that offer reliable and actionable insights into patient mortality risks.

Team

Project objective

In this project, we aim to accurately predict the APR Risk of Mortality using advanced machine learning models. This is crucial for effective clinical decision-making and optimizing healthcare resource management.


Understanding
the Dataset



Dataset Description

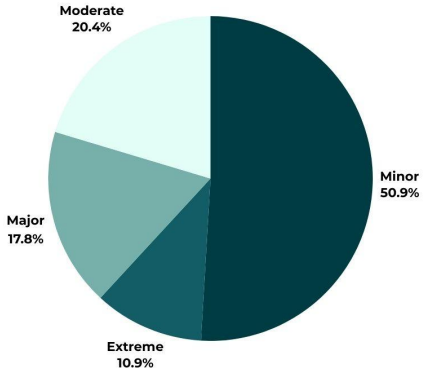
The dataset includes patient demographics, clinical diagnoses, procedures, and outcomes, providing a comprehensive base for predictive modeling.

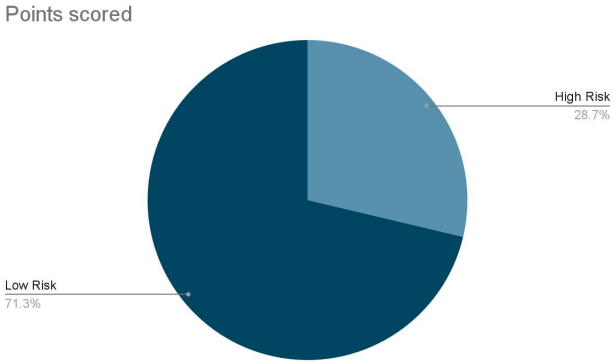
Data Source:

New York State Department of Health

https://health.data.ny.gov/Health/Hospital-Inpatient-Discharges-SPARCS-De-Identified/tg3i-cinn/about_data

	Hospital Service Area	Hospital County	Operating Certificate Number	Permanent Facility Id	Facility Name	Age Group	Zip Code - 3 digits	Gender	Race	Ethnicity	...	APR Severity of Illness Description	APR Risk of Mortality	APR Medical Surgical Description	Payment Typology 1	Payment Typology 2	Payment Typology 3	Birth Weight	Emergency Department Indicator	Total Charges	Total Costs
0	New York City	Bronx	7000006.0	1169.0	Montefiore Medical Center - Henry & Lucy Moses...	70 or Older	104	M	Other Race	Spanish/Hispanic	...	Major	Extreme	Medical	Medicare	Medicaid	NaN	NaN	Y	320922.43	60241.34
1	New York City	Bronx	7000006.0	1169.0	Montefiore Medical Center - Henry & Lucy Moses...	50 to 69	104	F	White	Not Span/Hispanic	...	Moderate	Minor	Medical	Private Health Insurance	NaN	NaN	NaN	Y	61665.22	9180.69
2	New York City	Bronx	7000006.0	1168.0	Montefiore Medical Center-Wakefield Hospital	18 to 29	104	F	Other Race	Spanish/Hispanic	...	Minor	Minor	Surgical	Medicaid	NaN	NaN	NaN	N	42705.34	11366.50
3	New York City	Bronx	7000006.0	3058.0	Montefiore Med Center - Jack D Weiler Hosp of ...	70 or Older	104	M	Other Race	Spanish/Hispanic	...	Major	Major	Medical	Medicare	Medicaid	NaN	NaN	Y	72700.17	12111.75





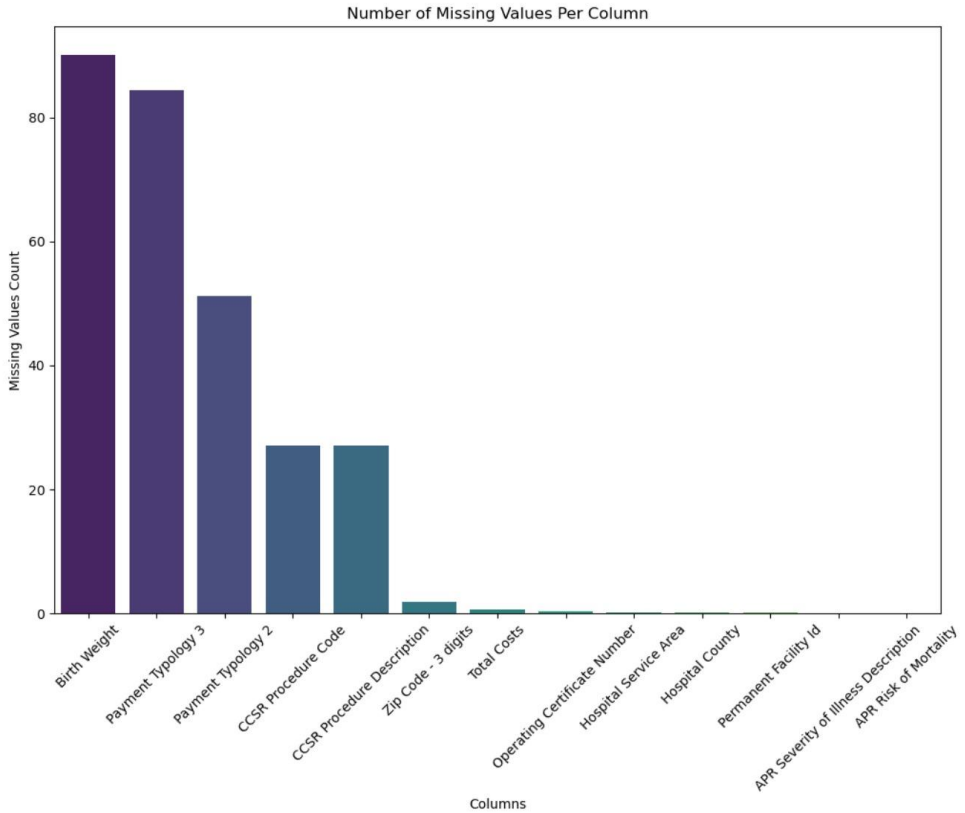
Target Column

APR Mortality

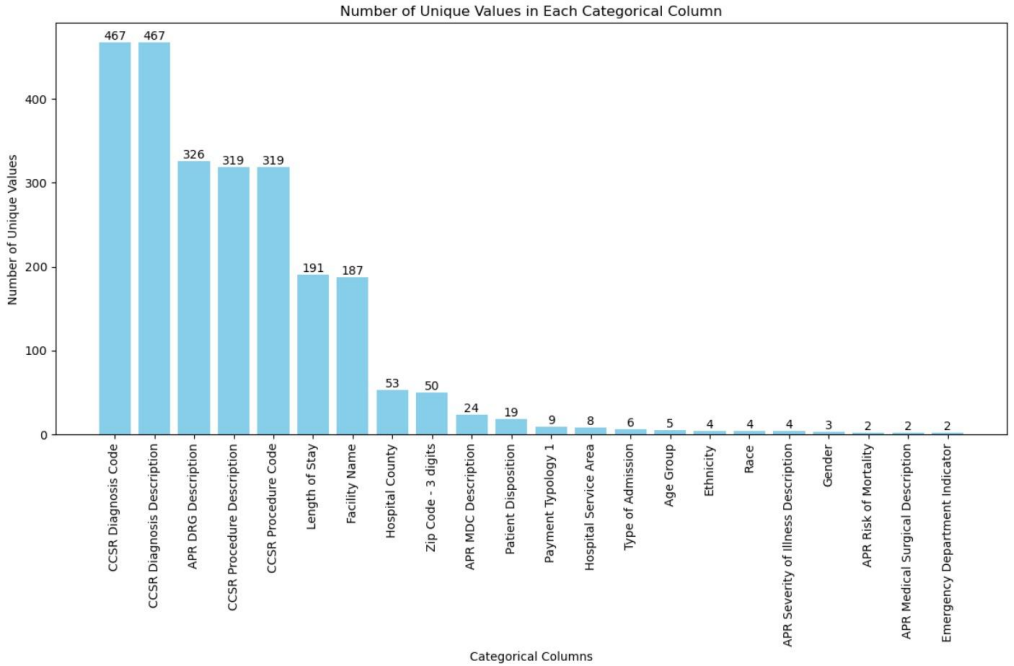
Value Counts:

Visualize the distribution of categories in the 'Target' column by counting the number of instances in each category.

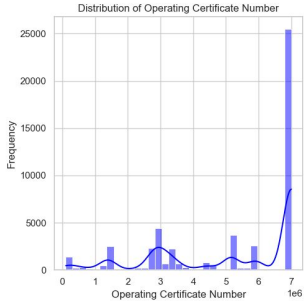
Null Values

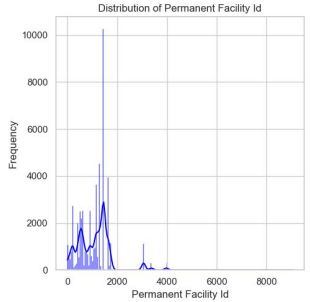

Number of Missing Values Per Column
Missing Values Count
0
20
40
60
80
Birth Weight
Payment Typology 3
Payment Typology 2
CCSR Procedure Code
CCSR Procedure Description
Zip Code - 3 digits
Total Costs
Operating Certificate Number
Hospital Service Area
Hospital County
Permanent Facility Id
APR Severity of Illness Description
APR Risk of Mortality
Columns

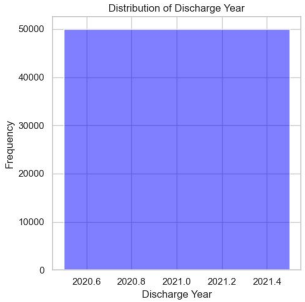
Unique Values of Categorical Columns

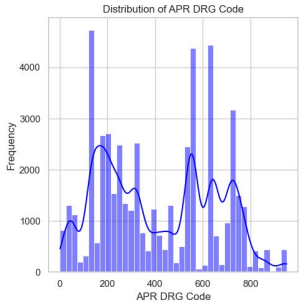


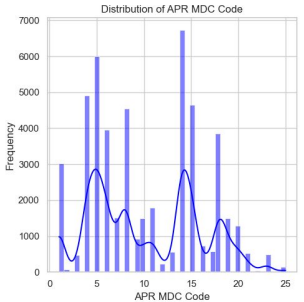
EDA

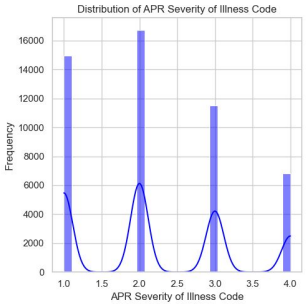

Distribution of Operating Certificate Number
Frequency
Operating Certificate Number
1e6

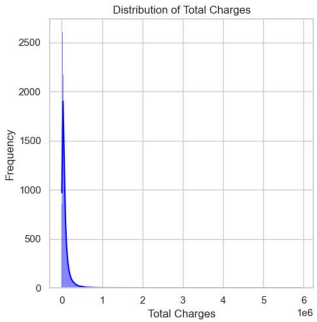

Distribution of Permanent Facility Id
Frequency
Permanent Facility Id

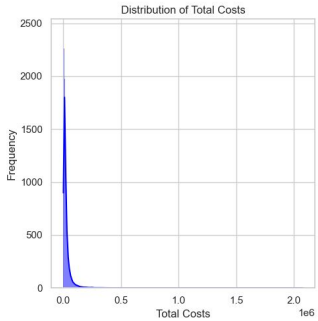

Distribution of Discharge Year
Frequency
Discharge Year


Distribution of APR DRG Code
Frequency
APR DRG Code


Distribution of APR MDC Code
Frequency
APR MDC Code

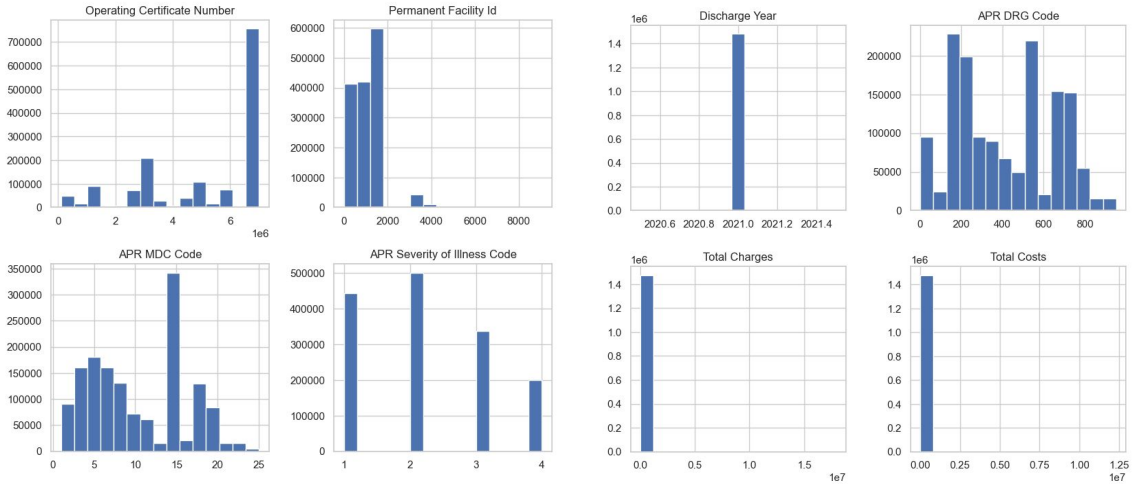

Distribution of APR Severity of Illness Code
Frequency
APR Severity of Illness Code


Distribution of Total Charges
Frequency
Total Charges
1e6

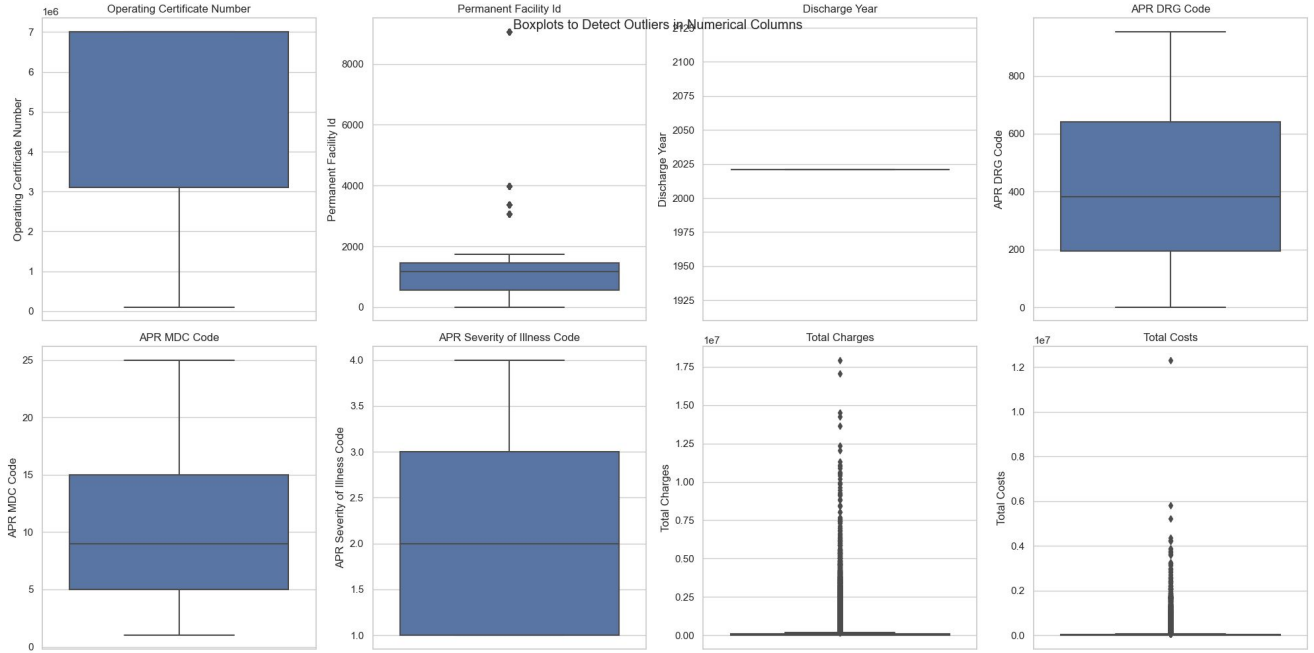

Distribution of Total Costs
Frequency
Total Costs
1e6

Numerical Columns

Distribution of Numerical Columns



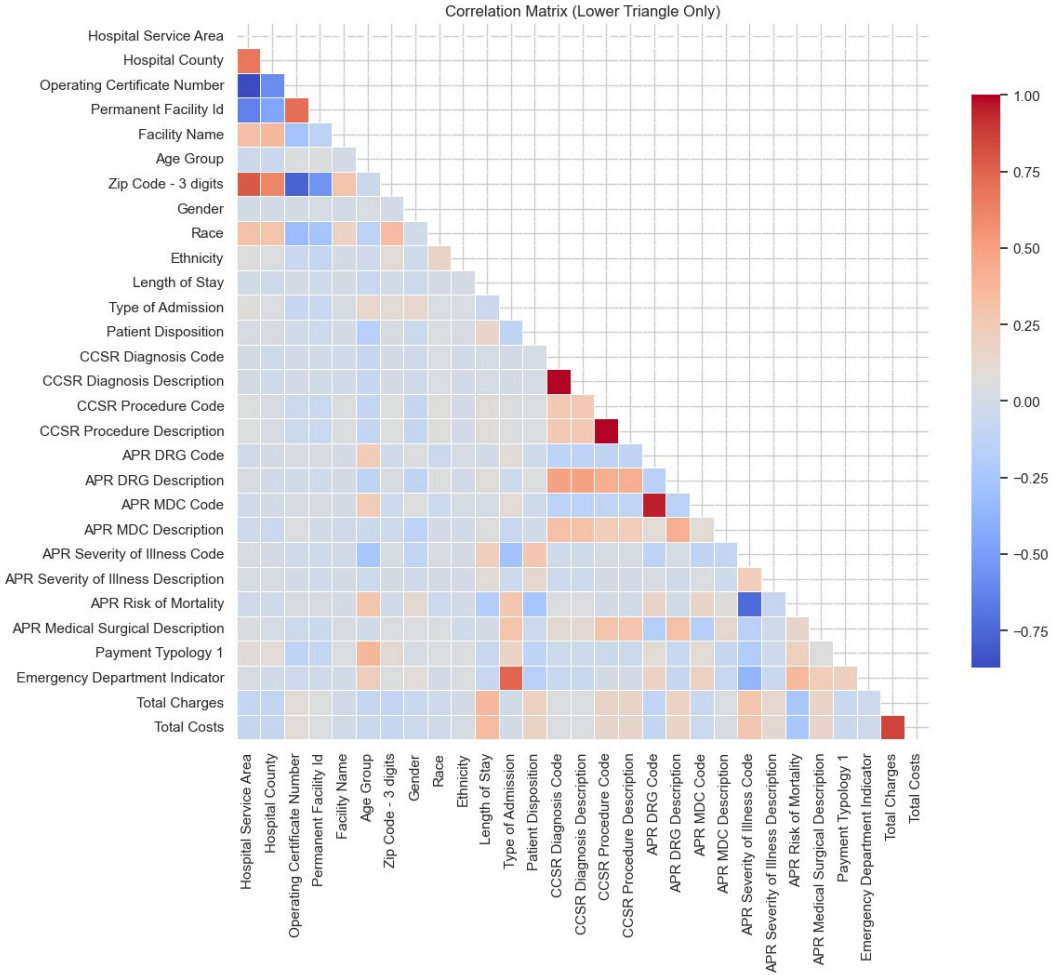
Outliers


Boxplots to Detect Outliers in Numerical Columns
Operating Certificate Number
1e6
Operating Certificate Number
Permanent Facility Id
Permanent Facility Id
Discharge Year
Discharge Year
APR DRG Code
APR DRG Code
APR MDC Code
APR MDC Code
APR Severity of Illness Code
APR Severity of Illness Code
Total Charges
1e7
Total Charges
Total Costs
1e7
Total Costs

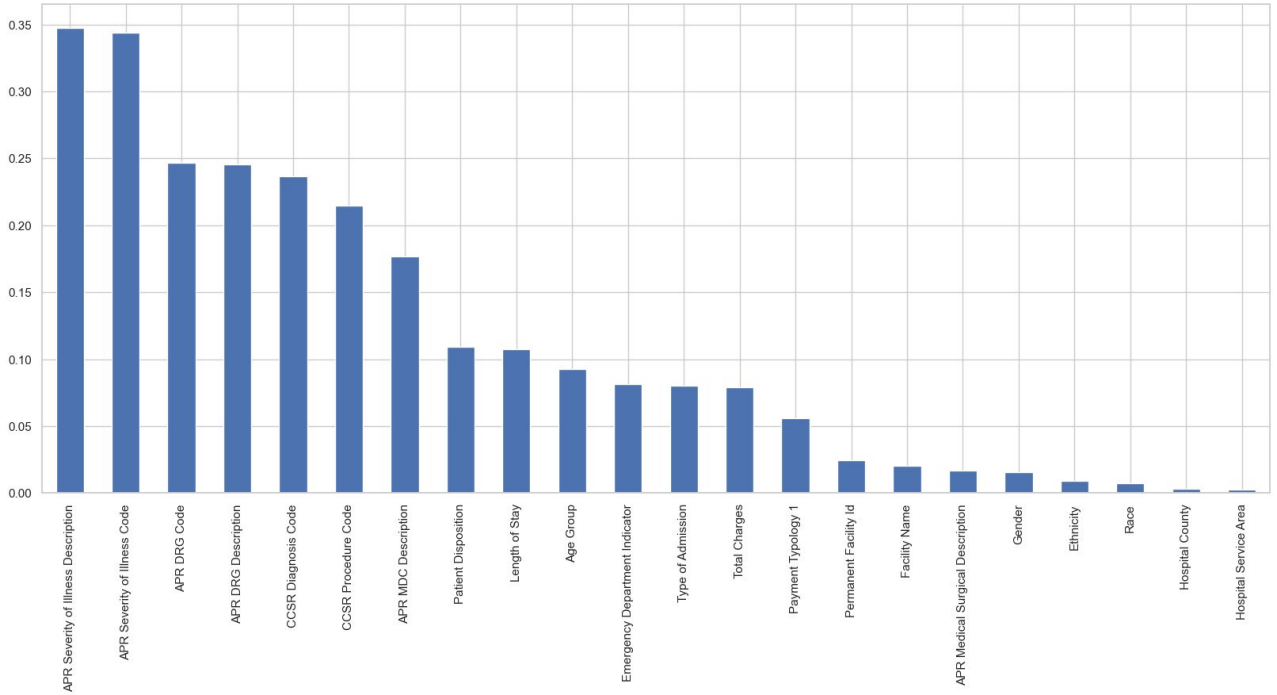
Correlation Matrix.

High Correlating Columns - Threshold(0.8)

Hospital Service Area	Operating Certificate Number
Operating Certificate Number	Permanent Facility Id
Hospital Service Area	Zip Code - 3 digits
Operating Certificate Number	Zip Code - 3 digits
CCSR Diagnosis Code	CCSR Diagnosis Description
CCSR Procedure Code	CCSR Procedure Description
APR DRG Code	APR MDC Code
APR Severity of Illness Code	APR Risk of Mortality
Type of Admission	Emergency Department Indicator
Total Charges	Total Costs



Feature Importance



Data Preprocessing

Categorizing the Data

During the first phase of preprocessing, we transformed the APR risk of mortality data, which originally had four categories, into a simpler binary format.

Treating Outliers

After confirming they were errors or extreme values that skewed our results, we removed these outliers to enhance the accuracy and reliability of our analyses.

Highly correlated data

Removing highly correlated features is crucial for reducing redundancy and preventing multicollinearity.

Handling Missing Values

We removed columns containing more than 20% null values, and in some cases, we filled in the missing values with the mode.

Managing Duplicates

Deleting duplicates is essential to minimize errors and ensure data integrity, leading to more accurate and reliable analyses by maintaining unique entries in the dataset.

Feature Selection

Selecting important features streamlines models by focusing on relevant variables, improving efficiency and predictive accuracy, and reducing complexity and overfitting.

Model Training

01 | Logistic Regression

02 | Random Forest

03 | Decision Tree

04 | XGBoost

05 | Support Vector Classifier

06 | Artificial Neural Network

LOGISTIC REGRESSION

Logistic Regression is used to predict the All Patient Refined (APR) Risk of Mortality from patient data. This model estimates the probability of mortality based on clinical and demographic features, categorizing each patient into 'low' or 'high' risk groups. It is valued for its interpretability, allowing healthcare providers to understand how different factors influence patient outcomes

Accuracy

89.70%

```
Logistic Regression Accuracy: 0.90%
Classification Report :
              precision    recall  f1-score   support

           0       0.83      0.83      0.83      2748
           1       0.92      0.93      0.92      6252

    accuracy                           0.90      9000
   macro avg       0.88      0.88      0.88      9000
weighted avg       0.89      0.90      0.89      9000
```

RANDOM FOREST

Accuracy

90%

Random Forest model is employed to classify patients into 'low' or 'high' mortality risk categories based on their clinical and demographic information. Random Forest is an ensemble learning method that builds multiple Decision Trees during training and outputs the class that is the majority vote of the individual trees. This approach effectively handles both linear and non-linear data, increases accuracy, and prevents overfitting.

```
The accuracy of the Random Forest model is  0.9075333333333333
Classification Report

              precision    recall  f1-score   support

           0       0.85      0.85      0.85      9271
           1       0.93      0.93      0.93     20729

    accuracy                           0.91     30000
   macro avg       0.89      0.89      0.89     30000
weighted avg       0.91      0.91      0.91     30000


Confusion Matrix:
[[ 7912  1359]
 [ 1415 19314]]
```

DECISION TREE

Accuracy

89.00%

Decision Tree model is utilized to classify patients into 'low' or 'high' risk of mortality based on their clinical and demographic data. Decision Trees work by splitting the data into subsets based on feature values, making decisions that aim to maximize class separation at each node. This results in a tree-like model of decisions, which is easy to interpret and understand, allowing clear visualization.

```
Decision Tree Accuracy: 0.89%
Classification Report :
              precision    recall  f1-score   support

           0       0.83      0.82      0.82      2748
           1       0.92      0.92      0.92      6252

    accuracy                           0.89      9000
   macro avg       0.87      0.87      0.87      9000
weighted avg       0.89      0.89      0.89      9000
```

X-GRADIENT BOOSTING

Accuracy

91%

Gradient Boosting model is used to predict the All Patient Refined (APR) Risk of Mortality based on patient data. Gradient Boosting constructs an ensemble of weak prediction models, typically Decision Trees, in a sequential manner where each subsequent model corrects the errors made by the previous ones. This method systematically improves predictions by focusing on difficult cases that earlier trees struggled with, enhancing accuracy, particularly in complex datasets with non-linear relationships.

XGBoost Accuracy: 0.91%

Classification Report :

	precision	recall	f1-score	support
0	0.86	0.84	0.85	2748
1	0.93	0.94	0.93	6252
accuracy			0.91	9000
macro avg	0.89	0.89	0.89	9000
weighted avg	0.91	0.91	0.91	9000

SUPPORT VECTOR CLASSIFIER

Accuracy

90%

the Support Vector Classifier (SVC) is used to classify patients into 'low' or 'high' mortality risk categories based on their clinical and demographic data. SVC works by finding the **optimal hyperplane** that best separates the data into two categories, effectively handling both linear and non-linear classification through the use of kernel tricks. This model is particularly effective in **high-dimensional spaces**

```
SVC Accuracy: 0.90%
Classification Report :
              precision    recall  f1-score   support

           0       0.82      0.86      0.84      2748
           1       0.94      0.92      0.93      6252

    accuracy                           0.90      9000
   macro avg       0.88      0.89      0.88      9000
weighted avg       0.90      0.90      0.90      9000
```

ARTIFICIAL NEURAL NETWORKS

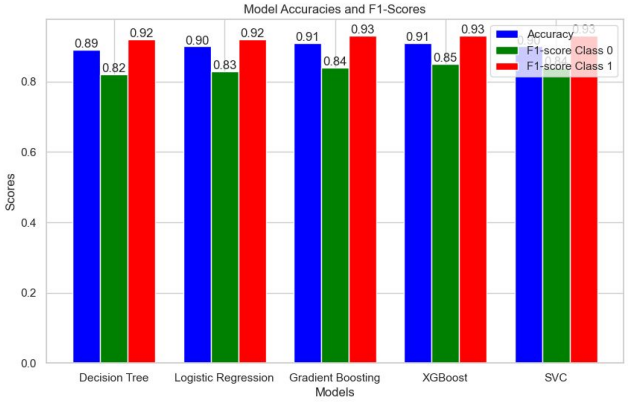
Accuracy

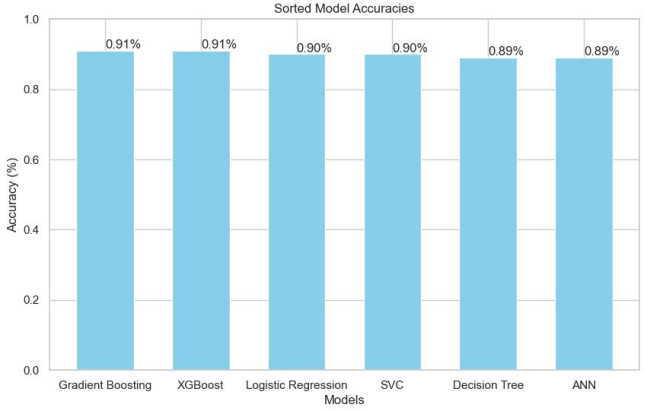
89%

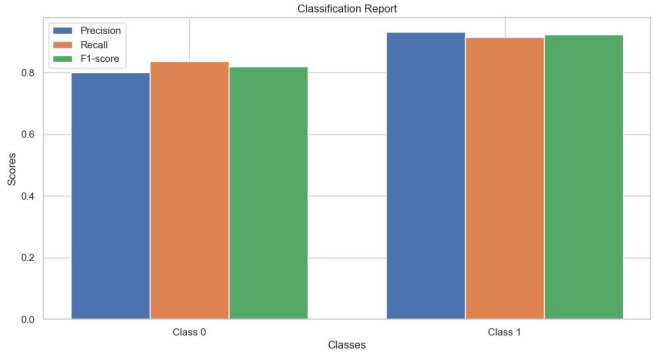
Artificial Neural Network (ANN) is applied to predict the All Patient Refined (APR) Risk of Mortality using patient data. This model leverages layers of interconnected neurons to process features, both clinical and demographic, to estimate the risk levels of mortality, categorizing outcomes into 'low' or 'high' risk. ANNs excel in handling non-linear relationships and complex interactions between variables, providing robust predictive performance even on imbalanced datasets.

```
282/282 ━━━━━━━━━━━━━━━━━━━━ 0s 639us/step
ANN Accuracy: 0.891777753829956
ANN Classification Report:
              precision    recall  f1-score   support

           0       0.82      0.86      0.84      2748
           1       0.94      0.92      0.93      6252

    accuracy                           0.90      9000
   macro avg       0.88      0.89      0.88      9000
weighted avg       0.90      0.90      0.90      9000
```

Model Comparison

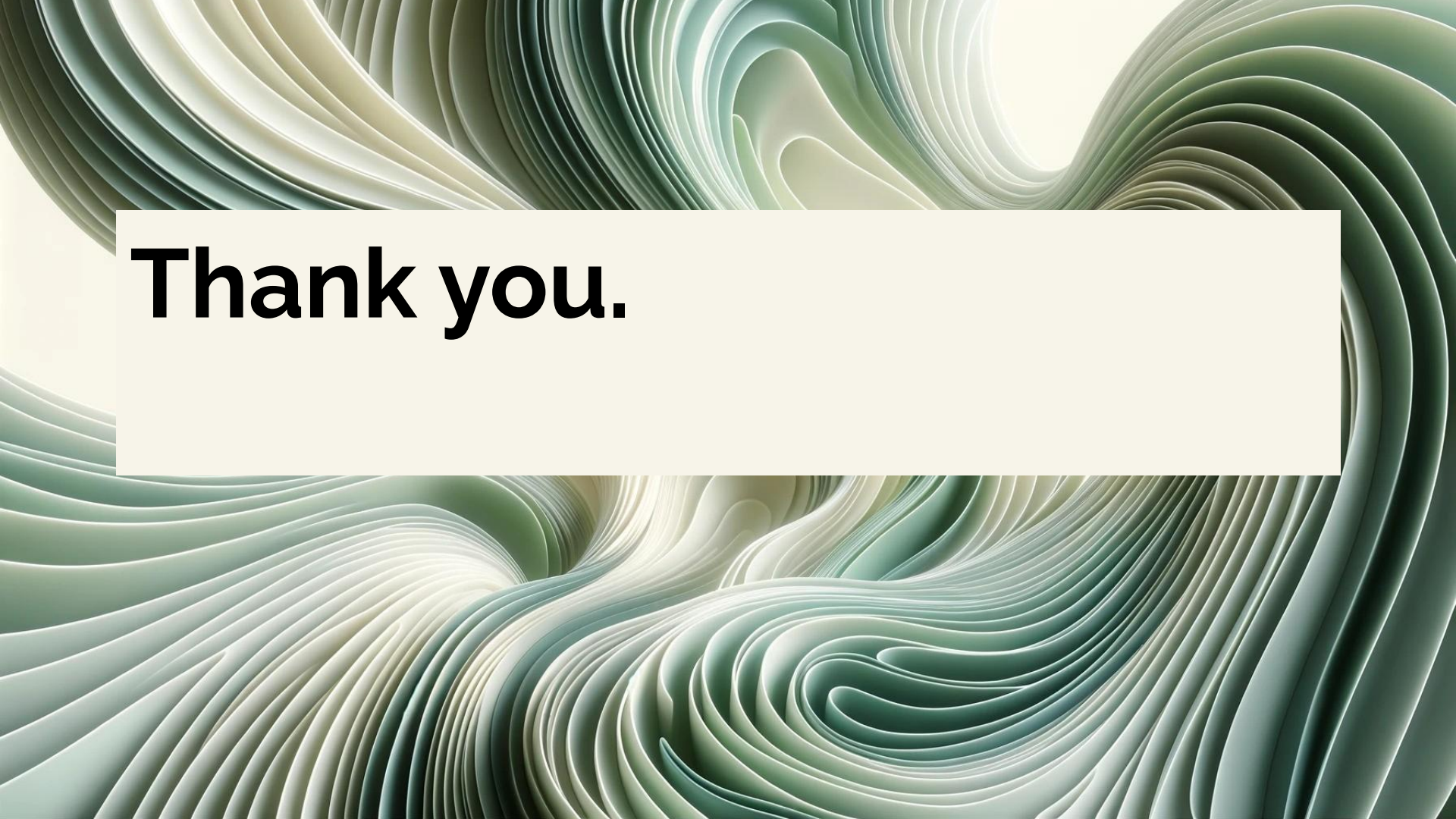






Proposed solution

XGBoost has demonstrated the highest accuracy among the models tested on our dataset, making it the preferred choice for predicting the **All Patient Refined (APR) Risk of Mortality**. Given its robust performance and efficiency in handling complex and diverse data, we are proceeding with **XGBoost** as our primary model for further development and application.


Thank you.